BAB II

TINJAUAN PUSTAKA

II.1. Machine Learning

Machine learning merupakan cabang ilmu bagian dari kecerdasan buatan (artificial intelligence) dengan pemrograman untuk memungkinkan komputer menjadi cerdas berperilaku seperti manusia dan dapat meningkatkan pemahamanya melalui pengalaman secara otomatis. Machine Learning memiliki fokus pada pengembangan sistem yang mampu belajar sendiri untuk memutuskan sesuatu tanpa harus berulangkali di program oleh manusia sehingga komputer menjadi semakin cerdas belajar dari pengalaman data yang dimiliki. Berdasarkan teknik pembelajarannya dapat dibedakan supervised learning menggunakan dataset (data traning) yang sudah berlabel sedangkan unsupervised learning menarik kesimpulan berdasarkan dataset (Endang Retnoningsih & Rully Pramudita 2020).

Algoritma yang termasuk kedalam teknik supervised learning diantaranya Decision Tree, K-Nearest Neighbor (KNN), Naïve Bayes, Regresi dan super Vector Machine (Fajar Sodik Pamungkas, Bayu Dwi Prasetya, Iqbal Kharisudin 2020). Beberapa Algoritma dalam unsupervised learning seperti Levenshtein Distance dan K-Means Clustering (Abdurahman et al.,2022).

Klasifikasi adalah cara yang dilakukan sebagai teknik untuk membentuk model klasifikasi dari contoh data pelatihan. Klasifikasi akan menganalisis input data dan membentuk model dengan menggambarkan kelas data. Dalam melakukan proses klasifikasi data mengacu pada metode kecerdasan buatan yang memfokuskan pada

pembelajaran mesin (machine learning). Banyak metode lain untuk mengetahui machine learning yang digunakan sebagai proses klasifikasi diantaranya K-Nearest Neighbor (K-NN) dan Naïve Bayes Classifier. Dengan menggunakan metode K-Nearst Neighbor (KNN) merupakan salah satu metode klasifikasi terhadap sekumpulan data yang berdasarkan mayoritas dari kategori dan tujuannya untuk mengklasifikasikan obyek baru berdasarkan atribut dan sample sample dari training data. Sehingga target output yang diinginkan mendekati ketepatan dalam melakukan pengujian pembelajaran (Permana Putra et al., 2022)

II.2. K-Nearest Neighbors

Algoritma *K-Nearest Neighbor* (KNN) merupakan algoritma klasifikasi jenis *supervised learning* yang mengklasifikasikan data dengan cara mengelompokkan data latih ke dalam kelompok-kelompok yang memiliki jarak terdekat dengan kelas target (Mardhyath et al., 2020). KNN merupakan metode yang paling banyak digunakan di dalam *data mining* dan *machine learning* karena implementasinya yang sederhana dan memberikan *performance* yang baik (Zhang et al., 2017). KNN terpilih sebagai salah satu dari 10 algoritma terbaik di dalam *data mining* dan *machine learning* (Wu et al., 2008). Prinsip dari KNN cukup sederhana, dimana dimulai dengan memilih jumlah K *neighbor*, kemudian lakukan perhitungan jarak, dan ambil *K* terdekat berdasarkan hasil perhitungan jarak. Kemudian, di antara K terdekat ini, hitunglah jumlah titik data dalam setiap kategori dan tempatkan titik data baru ke dalam kategori yang jumlah tetangganya maksimum (Maillo et al., 2017).

Terdapat 2(dua) isu utama di dalam KNN, yaitu terkait penentuan nilai *K* yang ideal dan metode perhitungan *distance* (Ertuğrul & Tağluk, 2017). Penentuan nilai *K*

sangat sensitif di dalam *KNN*. Nilai *K* yang terlalu kecil menyebabkan terlalu sensitif terhadap *noise* yang dapat berdampak pada terjadinya *overfitiing* dan nilai *K* yang terlalu besar dapat menyebabkan terjadinya *underfitting* (Beyer et al., 1999)(Subasi & Qureshi, 2023). Beberapa penelitian sudah membuktikan hal ini, dimana memiliki jumlah terdekat yang berbeda, diperoleh kinerja yang berbeda pula baik itu dari segi akurasi, *purity* maupun jumlah iterasi dari proses klasifikasi (Hidayati & Hermawan, 2021; Rao et al., 2022; Tulus & Situmorang, 2020).

Hal lain yang berpengaruh pada KNN selain jumlah *k* tetangga terdekat adalah metode untuk menghitung jarak terdekat *cluster,* yang juga dapat mempengaruhi kinerja algoritma K-NN. Secara *default* KNN memang menggunakan metode perhitungan *distance* dengan menggunakan *Euclidean Distance* (Santos et al., 2020). Metode perhitungan *distance* lain yang sering digunakan adalah *Manhattan Distance* (Han et al., 2012). Beberapa penelitian yang menggunakan metode *Euclidean* dan *Manhattan* menunjukkan adanya perubahan akurasi yang dihasilkan algoritma KNN bila dikaitkan dengan perhitungan *distance* pada proses klasifikasi (Hidayi & Hermawan, 2021; Iswanto et al.,2021; Margolang et al., 2022 Wahyono et al., 2020)

Untuk mengukur jarak antara data baru (testing) dengan data lama (training), pada algoritma K-NN dapat digunakan persamaan-persamaan pengukuran jarak sebagai berikut (Iswanto et al., 2021):

1. Eulidean Distance

$$D_{Euclidean} = ||x - y||_2 = \sqrt{\sum_{j=1}^{N} |x - y|^2} \quad \text{(II. 1)}$$

2. Minkowski Distance

$D_{Minkowski} = ||x - y||_{\lambda} = \sqrt[\lambda]{\sum_{j=1}^{N}|x - y|^{\lambda}}$ ..... (II. 2)

3. Manhattan Distance

$D_{Manhattan} = ||x - y||_{1} = \sum_{j=1}^{N}|x - y|$ ..... (II. 3)

4. Chebyshev Distance

$D_{Chebyshew} = ||x - y||_{\lambda} = \log \lambda - \infty \sqrt[\lambda]{\sum_{j=1}^{N}|x - y|^{\lambda}}$ ..... (II. 4)

Dalam algoritma K-NN selain nilai pemilihan metode pengukuran jarak, nilai K yang dipilih sebagai jumlah tetangga terdekat juga akan mempengaruhi akurasi dari klasifikasi karena dalam proses penentuan kelas data baru, algoritma ini menggunakan sistem voting (suara terbanyak) berdasarkan jumlah tetangga terdekat yang dimiliki oleh masing-masing data (Iswanto et al., 2021).

II.3. Klasifikasi Data

Klasifikasi data adalah proses pengurutan dan pengelompokan data ke dalam berbagai jenis, bentuk atau kelas berbeda lainnya. Klasifikasi data memungkinkan pemisahan dan klasifikasi data sesuai dengan persyaratan kumpulan data untuk berbagai tujuan bisnis atau pribadi. Ini terutama merupakan proses manajemen data. Analisis kelompok sebagai suatu metode untuk melakukan klasifikasi data menjadi beberapa kelompok dengan menggunakan metode pengukuran ukuran asosiasi, sehingga data yang sama berada dalam satu kelompok dan data yang memiliki perbedaan yang besar diletakkan dalam kelompok data lainnya.

Masukan untuk sistem analisis kelompok adalah sebuah data set dan kesamaan ukuran antara kedua data tersebut. Sedangkan hasil dari analisis kelompok adalah sejumlah kelompok yang membentuk sebuah partisi atau struktur partisi dari kumpulan

data dan deskripsi umum dari setiap kelompok, dimana hal ini sangat penting untuk analisis yang lebih dalam pada karakteristik yang terdapat pada data tersebut.

Pada proses klasifikasi, sebelum melakukan prediksi, perlu dilakukan proses pembelajaran terlebih dahulu. Proses pembelajaran tersebut memerlukan data. Data yang diperlukan pada saat proses klasifikasi terdiri atas dua jenis, yaitu (permana putra et al., 2022):

1. Data latih atau data training adalah data yang digunakan pada proses pembelajar dalam proses klasifikasi.
2. Data uji atau data testing adalah data yang digunakan pada proses prediksi dalam proses klasifikasi.

Langkah-langkah klasifikasi algoritma KNN adalah sebagai berikut (permana putra et al., 2022):

1. Tentukan parameter nilai K = banyaknya jumlah tetangga terdekat.
2. Hitung jarak antara data baru dengan semua data latih.
3. Urutkan jarak dan tetapkan tetangga terdekat berdasarkan jarak minimum ke K.
4. Periksa kelas dari tetangga terdekat.
5. Gunakan mayoritas sederhana dari kelas tetangga terdekat sebagai nilai prediksi data baru.

II.4. Mangga

Mangga adalah salah satu jenis buah yang unggul dan sangat digemari untuk dikomsumsi oleh masyarakat, buah manga sendiri banyak digemari masyarakat karena rasannya bervariasi dengan banyaknya jenis buah mangga tersebut bagi orang awam akan

merasa kesulitan dalam klasifikasi jenis-jenis buah mangga. Mangga yang sering kali dikomsumsi oleh kebanyakan masyarakat yaitu jenis mangga harum manis, mangga manalagi dan mangga apel maka daripada itu dibutuhkanlah suatu metode yang cocok untuk menyelesaikan permasalahan penelitian ini.

II.5. Pengenalan Pola

Pengenalan pola merupakan serangkaian proses yang berkesinambungan, dimulai dari proses deteksi/segmentasi, sistem ekstraksi, dan proses pengukuran kemiripan atau proses pengukuran kesamaan atau proses pengenalan. Komputer kini dapat digunakan untuk mengenali masukan berupa pola yang digunakan sebagai informasi. Pengenalan pola juga dapat mengklasifikasikan, namun tidak semua algoritma klasifikasi dapat melakukan pengenalan pola. Algoritma K-Neareast Neighbor memiliki keunggulan dalam proteksi terhadap noise.

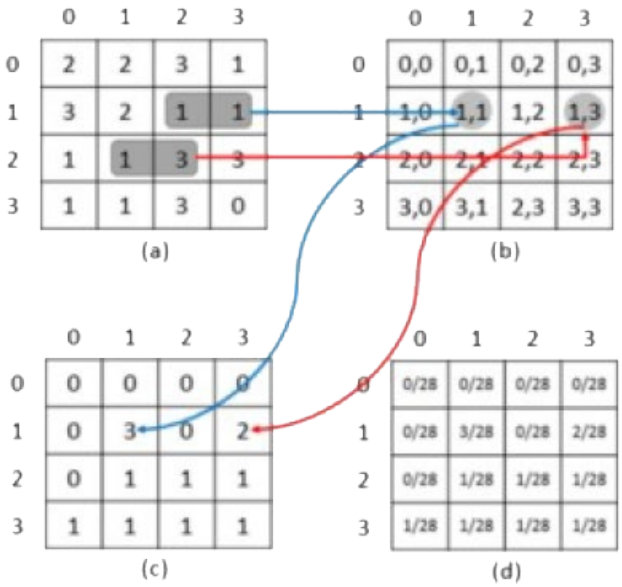
II.6. Pengolahan Citra Digital

Pengolahan citra digital merupakan ilmu yang mempelajari permasalahan yang berkaitan dengan peningkatan kualitas citra (peningkatan kontras, konversi warna, restorasi citra), transformasi citra (rotasi, translasi, penskalaan, transformasi geometri), hingga pemilihan fitur citra (image feature) yang optimal untuk tujuan analisis, melakukan proses pengambilan informasi atau mendeskripsikan objek atau mengidentifikasi objek pada gambar, melakukan kompresi atau pengurangan data untuk keperluan penyimpanan data, transmisi data, dan waktu pemrosesan data. Masukan dari pengolahan citra adalah gambar sedangkan keluarannya adalah gambar yang diperoleh dari pengolahannya.

II.7. Ekstraksi Ciri GLCM

Pada penelitian ini menggunakan ekstraksi ciri tekstur GLCM. GLCM adalah analisis tekstur yang sudah banyak digunakan dan hasil yang diperoleh dari matriks co-occurrence lebih baik dari metode diskriminasi tekstur lainnya. GLCM menghitung fitur statistic berdasarkan tingkat keabuan (grayscale) gambar.

GLCM adalah matriks dimana jumlah baris dan kolom sama dengan jumlah tingkat keabuan, G dalam image. Elemen matriks P($i,j|\Delta x$, Δy) adalah frekuensi relative yang dipisahkan oleh jarak-jarak pixel (Δx, Δy) Elemen matriks juga direpresentasikan sebagai P(i,j|d, Θ) yang berisi nilai probabilitas orde kedua untuk perubahan antara tingkat keabuan i dan j pada jarak d dan sudut Θ tertentu. Berbagai fitur diekstraksi GLCM, G adalah jumlah tingkat keabuan yang digunakan dan μ adalah rata-rata dari P, μx, μy, δx dan δy adalah rata-rata dan simpangan baku dari Px dan Py. Adapun matrik dari GLCM dapat dilihat pada gambar II.1



Gambar II.1. GLCM Matrix (a) Matrix Citra Asli (b) GLCM Index (c) Co-Occurrence Matrix (d) Probability

$$Px(i) = \sum_{i=0}^{G-1} P(i,j) \text{ dan } Py(j) = \sum_{i=0}^{G-1} P(i,j) \quad (1)$$

$$\mu x = \sum_{i=0}^{G-1} iPx(i) \text{ dan } \sum_{j=0}^{G-1} jPyj \quad (2)$$

$$\delta_x^2 = \sigma_{i=0}^{G-1} ((\text{Px}(i)) \ \text{-}\mu_x(i))^2 \quad (3)$$

$$\delta_y^2 = \sigma_{j=0}^{G-1} ((\text{Py}(j)) \ \text{-}\mu_y(j))^2 \quad (4)$$

GLCM memiliki fitur sebanyak 12 fitur. Pada penelitian ini hanya mengambil 4 fitur yaitu Homogeneity, Energy, Contrast, dan Correlation.

a. Homogeneity

Homogeneity adalah nilai yang mengukur kedekatan distribusi elemen dalam GLCM dengan diagonal GLCM.

$$\text{Homogeneity} = \sum_{i=0}^{G-1} \quad \sum_{j=0}^{G-1} \frac{1}{1+(i-j)^2} P(i,j) \quad (5)$$

b. Energy

Energy adalah jumlah elemen kuadrat GLCM yang di normalisasi.

$$\text{Energy} = ASM = \sum_{i=0}^{G-1} \quad \sum_{j=0}^{G-1} P(i,j)^2 \quad (6)$$

c. Contrast

Contrast adalah untuk mengukur intensitas kontras di antara piksel tetangganya.

$$Contrast = \sum_{n=0}^{G-1} n^2 \{\sum_{i=1}^{G} \quad \sum_{j=1}^{G} P(i,j)\}, |i-j|| = n \quad (7)$$

d. Correlation

Correlation adalah untuk mengukur seberapa berkorelasi piksel dengan tetangganya.

$$\sum_{i,j=0}^{level-1} Pi,j \left[\frac{(i-\mu i)(j-\mu j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right] \quad (8)$$

II.8. Confusion Matrix

Confusion Matrix merupakan salah satu metode yang dapat digunakan untuk mengukur performansi suatu metode klasifikasi. Pada dasarnya confusion matrix ini berisi informasi membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi sebagaimana mestinya. Saat mengukur kinerja menggunakan confusion matrix ada 4 (empat) istilah yang menggambarkan hasil dari proses klasifikasi. Keempat istilah tersebut adalah True Positive, True Negative, False Positive dan False Negative (Muhammad Ali Imron et al., 2020).

Tabel II.1. Confusion Matrix

Kelas	Terklarifikasi Positif	Terklarifikasi Negatif
Positif	TP (True Positive)	FN (False Negative)
Negatif	FP (False Positive)	TN (True Negative)

Dimana:

1. True Positive, jumlah data positif yang diklasifikasikan dengan benar oleh sistem.
2. True Negative, jumlah data negatif yang diklasifikasikan dengan benar oleh sistem.
3. False Negative, jumlah data negative namun diklasifikasikan salah oleh sistem.
4. False Positive, jumlah data positif namun diklasifikasikan salah oleh sistem.

Berdasarkan nilai True Negative (TN), False Positive (FP), False Negative (FN), dan True Positive (TP) dapat diperoleh nilai akurasi, presisi dan recall. Nilai akurasi menggambarkan seberapa akurat sistem dapat diklasifikasikan data secara benar. Dengan kata lain, nilai akurasi dapat diperoleh dengan persamaan (6). Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. Presisi dapat diperoleh dengan persamaan (7)

sementara itu, recall menunjukkan berapa persen data kategori positif yang terklasifikasikan dengan benar oleh sistem. Nilai recall diperoleh dengan persamaan (8)

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (6)$$

$$\text{Presisi} = \frac{TP}{FP+TP} * 100\% \quad (7)$$

$$\text{Recall} = \frac{TP}{FN+TP} * 100\% \quad (8)$$

II.9. Penelitian Terkait

Dalam penelitian ini, digunakan artikel-artikel penelitian yang digunakan sebagai bahan referensi yang dirangkum di dalam tabel-tabel. Tabel II.2 Penelitian Terdahulu.

Tabel II.2 Penelitian Terdahulu

Penulis	Judul	Tahun	Persamaan	Perbedaan
Fathur Rizal, Nadiyah	Perbandingan Algoritma K-Nearest Neighbor Untuk Klasifikasi Jenis Mangga Menggunakan Berdasarkan Fitur Gray Level Co-Occurrence Matric dan Fitur Warna	2021	Sama-sama menggunakan pengukuran jarak Eulidean Distance	Tidak menggunakan pengukuran jarak Manhattan Distance
Zhou Chen, Lan Jiang Zhou, Xuan Da Li, Jia Nan Zhang, Wen Jie Huo	The Lao Text Classfication Method Based on KNN	2020	Sama-sama menggunakan metode KNN dan pengukuran jarak Minkowski Distance dan Euclidean Distance	Tidak menggunakan pengukuran jarak Manhattan Distance
T.Srinivas Rao, M,Hema, K.Sai Priya, K.Vamsi Krishna, M.Sakhavath Ali	Iris Flower Classfication Using Machine Learnig	2021	Sama-sama menggunakan pengukuran jarak Euclidean Distance dan Manhattan Distance	Tidak membahas penentuan nilai K

Purwa Hasan Putra, Muhammad Syahputra Novelan, Muhammad Rizki	Analysis K-Nearest Neighbor Method In Classification Of Vegetable Quality Based On Color	2022	Sama-sama menggunakan metode KNN	Tidak membahas pengukuran jarak Euclidean Distance dan Manhattan Distance
Wenchao, Yilin Bei	Medical Health big Data Classification on KNN Classfication Algorithm	2020	Sama-sama menggunakan metode KNN	Tidak membahas pengukuran jarak Euclidean Distance dan Manhattan Distance
Nunsina, Zakarias Situmorang	Analiysis Optimization K-Nearest Neighbor Algorithm with Certainty Factor in Determining Student Career	2020	Sama-sama menggunakan algoritma KNN	Tidak membahas pengukuran jarak Euclidean Distance dan Manhattan Distance
Soo-Tai Nam, Seong-Yoon Shin, Chan-Yong Jin	A Reconstruction of Classification for Iris Species Using Euclidean Distance Based on a Machine Learning	2020	Sama-sama algoritma KNN dan pengukuran jarak Euclidean Distance	Tidak menggunakan pengukuran jarak Manhattan Distance

Frencis Matheos Sarimole, Anita Rosiana	Classification of Maturity Levels in Area Fruit Based on HSV Image Using The KNN Method	2022	Sama-sama menggunakan metode KNN	Tidak membahas pengukuran jarak Euclidean Distance dan Manhattan Distance
Lidya Ningsih, Putri Cholidhazia	Classification of Tomato Maturity Levels Based on RGB and HSV Color Using KNN Algorithm	2022	Sama-sama menggunakan algoritma KNN dan Pengukuran jarak Euclidean Distance	Tidak menggunakan pengukuran jarak Manhattan Distance
Siddharth Nandakumar Chikalkar	K-Nearest Neigbors Machine	2020	Sama-sama menggunakan metode KNN dan pengukuran jarak Euclidean Distance dan Manhattan Distance	Tidak menggunakan pengukuran jarak Manhattan Distance

II.10. Kontribusi Penelitian

Penelitian yang dilakukan memiliki beberapa kontribusi yaitu:

1. Mengetahui pengaruh nilai K dan pengukuran jarak eulidean distance dan manhattan distance dalam klasifikasi jenis buah mangga yang dibangun menggunakan metode K-Nearest Neighbors.
2. Menghasilkan tingkat akurasi yang paling optimal antara pengukuran jarak eulidean distance dan manhattan distance dalam metode K-Nearest Neighbors.